Automation and Robotics in High Throughput Protein Crystallography

By Bernhard Rupp, Macromolecular Crystallography and Structural Genomics Group of the Lawrence Livermore National Laboratory, University of California

Dr Bernhard Rupp heads the Macromolecular Crystallography and Structural Genomics Group at the Lawrence Livermore National Laboratory, University of California, US. He also established the high-throughput crystallisation facility of the TB Structural Consortium, one of the nine NIH-NIGMS funded Protein Structure Initiatives. With a broad background in instrumentation, crystallography and structural chemistry and biology, Dr Rupp has held research positions in the US, Austria, Germany, Switzerland and Israel. In addition, he has conducted work on neurotoxins, superantigens and mycobacterium tuberculosis drug target structures. Dr Rupp is Adjunct Professor for Molecular Structural Biology at the University of Vienna, Austria.

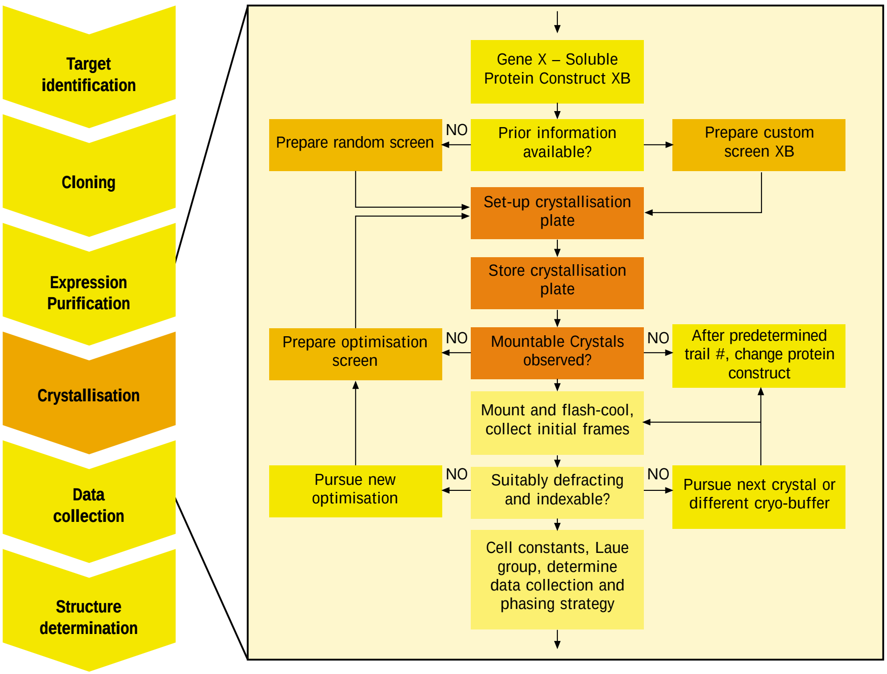
The availability of high resolution crystal structures is of key importance for structure guided drug design. Publicly and commercially funded structural proteomics efforts have led to accelerated technical development and increased availability of high throughput robotics for protein crystallography, extending from parallel protein expression to fully automated protein crystallisation robotics and automated sample mounting, data collection and structure solution methods. Successful implementation of laboratory automation requires careful planning, with a balance of expectations and resources versus long-term costs and return on investment. Not every research environment requires the same type or degree of automation, and operations research and process analysis are necessary to select the robotic automation layout most suitable to achieve maximal overall efficiency. Based on process outlines and a review of critical steps, we provide examples for cost-effective, modular designs for crystallography laboratory automation, emphasising the importance of automated process scheduling, data capture and adequate database support.

Rapid technical developments in automated high throughput crystallisation have led to the widespread availability of relatively affordable robotic automation. The NIH Protein Structure Initiative, PSI-I, (1) provides significant public funding to nine P50 structural genomics (SG) centres in the US, and similar initiatives in Europe (SPINE), Canada, Japan and Israel have been funded. A main objective of these centres is the advancement of high-throughput crystallography. Given the potentially enormous rewards of structure guided drug development (2-5), it comes as no surprise that a substantial number of biotech ventures were able to attract capital to develop and implement advanced custom robotic automation in protein crystallography. Consequently, many vendors have now realised the market potential for automation of crystallisation, and are competing with a wide palette of products ranging from basic liquid handling robots to integrated high throughput crystallisation workstations, imaging stations and crystal mounting robots.

This poses the question of how to spend available funds most wisely to maximise return on the investment into robotics. Significant differences in objectives, throughput goals, capital, organisational structure and talent pool exist between academic laboratories and large scale pharmaceutical drug discovery efforts, and each requires a specific implementation plan. On an industrial scale, where the high-throughput environment can process hundreds of crystals a week, it may be neither necessary nor efficient to pursue every recalcitrant target to completion, while in a small scale academic setting, careers can depend on the determination of one specific structure.

Figure 1: Schematic Flow of a Crystallographic Structure Determination (Left)

Right panel insert shows a more detailed view of basic crystallisation tasks and their relation. Shading indicates basic liquid handling (medium orange), crystallisation plate set-up and handling (dark orange), and mounting and data collation tasks (light yellow). Absent from the flow diagrams are the feedback mechanisms from structure analysis to target modification as well as iterative ligand screening and optimisation.


Target identification
Cloning
Expression Purification
Crystallisation
Data collection
Structure determination
Gene X – Soluble Protein Construct XB
Prepare random screen
NO
Prior information available?
Prepare custom screen XB
Set-up crystallisation plate
Store crystallisation plate
Prepare optimisation screen
NO
Mountable Crystals observed?
NO
After predetermined trail #, change protein construct
Mount and flash-cool, collect initial frames
Pursue new optimisation
NO
Suitably defracting and indexable?
NO
Pursue next crystal or different cryo-buffer
Cell constants, Laue group, determine data collection and phasing strategy

ROBOTIC TECHNOLOGY IN THE CRYSTALLOGRAPHY LABORATORY

The process of crystallographic structure determination can be broken up into a number of successive task blocks (see Figure 1), beginning with target selection, followed by protein production, crystallisation, data collection and finally structure determination. The highest demand for automation is generally found in screening steps, where repetitive tasks of modest complexity can be conveniently handled by robotics. Protein crystallisation is currently the most prominent candidate for full automation, while at the same time the front-end of protein production begins to undergo a similar transformation with increasing use of small scale, high throughput parallel expression and purification screening techniques. Novel plasmid vectors, expression autoinduction, affinity-tagged proteases of high specificity, and modular parallel chromatography equipment have contributed major advances towards automation in protein production. On the other hand, structure solution, model building and refinement are conducted entirely *in silico* and, given the rapid development of automated computational crystallography (3, 6-8), are generally not considered throughput limiting factors.

Success statistics compiled from the NIH PSI-I initiatives provide a baseline for the expected attrition from gene targeting up to the availability of a validated crystal structure. The trend is sobering and reveals substantial attrition: on average, slightly more than a third of soluble proteins form crystals, and about another 30-50 per cent of those crystals actually yield data leading to structures. Together, the probability of obtaining a structure from a purified protein is about 10-20 per cent. Losses from construct to purified protein are equally high (in particular for other than soluble, bacterial proteins), and average success rates are again in the order of about 20-30 per cent. Clearly, automation of protein production is crucial to the success of a structural proteomics pipeline, as emphasised by the increasing importance of exploring multiple constructs, orthologs (9), or engineered proteins (10-12) in order to obtain crystallisable proteins. Serious late stage losses finally occur in the step from harvestable crystal to diffraction data, caused by handling of the fragile crystals during harvesting, cryo-protection, mounting and annealing (13), and through radiation damage [14]. Optimisation of the post-crystallisation handling is thus equally important to achieve high overall process efficiency.

The overall objective of automation is an increase in the efficiency of the entire process, with specific implementation examples of liquid handling, cocktail preparation, crystallisation plate set-up, plate manipulation, observation of crystal growth and crystal mounting. A crucial element in the evaluation of throughput expectations is to define current, and anticipate future, rate limiting steps. A singular deployment of high speed robotics (in particular if intended to significantly increase throughput), will likely create new bottlenecks downstream: given sufficient protein supply, even with affordable equipment, about 20-30 96-well crystallisation trays per eight hour shift can be set up with ease. Assuming a plate shelf life time of four months and eight observations in that period, roughly 2,400 plates accumulate and require an observation roughly every two seconds. Cleary, image acquisition and analysis of outcomes need to be highly automated, and corresponding action in response to the outcome, such as scheduling of mounting or automated set-up of optimisation and harvesting plates, needs to follow in a timely manner. Without paralleling measures, increased robotic protein crystallisation throughput also tends to outstrip the protein production capacity (15), providing an example for upstream shift of a rate limiting process step. High crystallisation throughput of even a few proteins per day is not sustainable without careful planning of the entire automation set-up and thorough operations review.

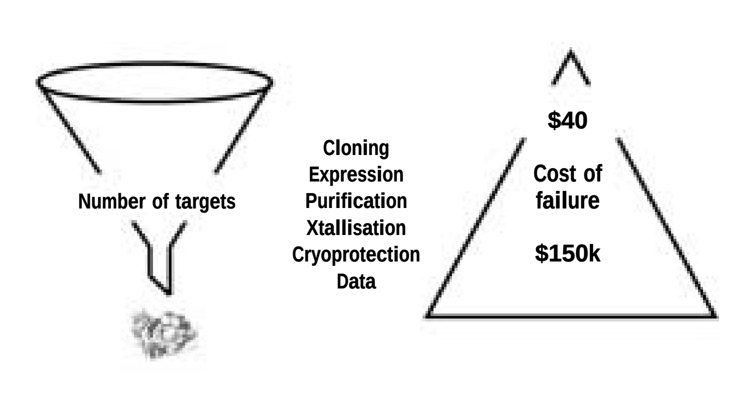
One of the biggest advantages of automation is miniaturisation, allowing comprehensive parallel screening of large sample sets (multiple constructs, orthologs, media and so on) with very little material. As a consequence, statistically sound go/no-go decisions can be made early in each successive screening step, and the pursuit of a target already showing warning signs of limited likelihood of success can be avoided. Early go/no-go decisions are common practice in the pharmaceutical industry, and academic examples are the ‘two-tiered’ approaches to crystallisation (16, 17), or the estimate of a point of diminishing return in random screening based on statistical analysis (18). Failure in later stages of the process, as exemplified by losing harvestable crystals or the failure of new therapeutic drug candidates in late phases of clinical trials due to unexpected drug interactions (19), tends to become increasingly costly (see Figure 2, see page 56).

A somewhat underestimated consequence of automation is the rapid generation of data during multiple levels of successively branched screening steps (such as expression, solubilisation, crystallisation, cryo-protection) (21). Despite considerable sample attrition, the amount of data generated at each screening step can rapidly outstrip the capability to analyse them (22). Consequently, even for a small effort, the capturing of primary data at the source directly into a relational database via automated scripts or a laboratory information management system (LIMS) is important. Elaborate LIMS packages are mandatory for the pharmaceutical and biotech industry (23), largely due to regulatory requirements (21 CFR part 11 compliance), but smaller and flexible systems are

Figure 2: Target Attrition Versus Cost of Failure

The later a target is abandoned in the process, the more costly steps it has passed through and the more resources it has consumed. As extreme examples, a failure to clone a target out of DNA in high throughput mode is probably limited to waste of a primer pair, some chemicals and a small amount of labour. The other extreme would be a dataset that cannot be phased. At this point, the target has passed all cost-intensive experimental stages and accumulated maximal value. Similarly, losing crystals during cryo-soaking is very expensive, affirming that more effort should be spent on systematic investigation of cryo-protection (20). Estimated total cost per publicly-funded structure decreased from $500k/S in 2000 to $150k/S in late 2003 and is expected to decrease further to about $50k/S in 2005, when the initial phase of the PSI will end.

Source: NIH report, Planning for the Protein Structure Initiative, *Sept 2003*



becoming available for proteomics and crystallography (24-26) and some are being developed by the PSI centres (27). In addition, with increasing automation, tightly integrated process control, multiple feedback steps, real-time data processing and decision-making, and machine learning for predictive purposes (21) are becoming a major component of any automation intensive laboratory. Although initial cost and effort to implement a LIMS are not insignificant, the return on investment (ROI) can be less than a year (28). A well-designed data repository based on a relational database will support access to existing public databases, allowing cross-database queries of supporting information that can be used at various levels in the decision-making process (29); at the top level the need to construct new ontologies for data description and data mining of complex knowledge (30) need to be anticipated.

Full walk-away automation up to, but not including, harvesting is conceivable given the current equipment available on the market, and has been demonstrated (albeit at substantial cost) in several custom-made industrial designs (4, 31). Cocktail preparation, plate set-up, automated crystal recognition and subsequent optimisation can be integrated with plate handling robotics and provide no principal (nonetheless financial) challenges. Process automation currently stops at the harvesting stage, largely due to the expense of micromanipulation, and the need for advanced machine vision tools to allow real-time processing of the events during crystal harvesting. However, new reproducibly manufactured mounting loop designs (32, 33) and micro-manipulation actuators for robots are being developed, and will eventually address this remaining manual bottleneck. Once the crystals are safely cryo-protected, robotic mounting of sample pins has become standard on HTPX synchrotron beam lines and in larger biotech companies and laboratories (34-38).

An issue that usually affects commercial ventures more than academics is the need for licensing of patented materials, processes or copyrighted code. In the US, for example, the use of nanoliter drop technology in crystallisation is protected and was the subject of an infringement dispute (US Patent 6,296,673). Given the ever-increasing intertwinement of academics and commercial contract work, patent and licensing issues cannot be ignored.

Vendor or third-party on-site service contracts constitute a nearly mandatory, but easily forgotten expense for complex equipment. Particularly if no dedicated local support or engineering team is available, service contracts can provide a good return on investment. On-site service contracts in general seem to cost about 10 per cent of the equipment price per year, and their value for any production schedule is evident.

CONCLUSION

Automation in protein crystallography is becoming increasingly attractive, and for successful deployment in any laboratory setting, the expected benefits must be offset against the total cost of ownership. Although basic equipment is affordable, there is more to automation than just purchasing a liquid handler and/or plate incubator with an integrated camera. Automation for high throughput should parallel a change in mindset, including a re-analysis of the whole process of crystallisation and protein production. Robotic high throughput crystallisation screening in micro-drops allows, in an unprecedented way, the implementation of strategies that identify lead candidates early, and provides the opportunity for corrective action at the protein construct level – before more resources are expended on low probability targets likely to fail in later steps.

High-throughput crystallisation of even a few proteins per day is not sustainable without careful planning of the entire automation set-up and a thorough operations review. A singular deployment of high speed robotics, in particular if intended to significantly increase throughput, will likely create new bottlenecks downstream. Data capture, warehousing and curating is of paramount importance, not just for process control, but also for successful data mining of highly dimensional and complex proteomics data. If proteomics and crystallisation data analysis is to evolve from basic frequency and propensity analysis to truly predictive models of specific inference, it is mandatory that common metrics for scores and minimal standards for experimental design are maintained. Realistic planning and full consideration of high throughput process and design principles will go a long way to accomplish a successful and financially sound transition into robotic high-throughput crystallography. ◆

Acknowledgements

Li Chen (RCSB Rutgers) has helped in extracting information from the PSI target database. LLNL is operated by the University of California for the US DOE under contract

W-7405-ENG-48. This work was funded by NIH P50 GM62410 (TB Structural Genomics) centre grant.

The author can be contacted at br@llnl.gov

References

1. Norvell JC and Zapp-Machalek A, Structural genomics programs at the US National Institute of General Medical Sciences, *Nature Struct Biol Suppl*, 7: p931, 2000
2. Harris T, The commercial use of structural genomics, *Drug Discovery Today*, 6(22): p1,148, 2001
3. Blundell TL, Jhoti H and Abell C, High-Throughput Crystallography for Lead Discovery in Drug Design, *Nature Reviews Drug Discovery*, 1: pp45-54, 2001
4. Burley S, The FAST and the curios, *Modern Drug Discovery*, 7(5): pp53-56, 2004
5. Yon J and Jhoti H, High-throughput structural genomics and proteomics: where are we now?, *Targets*, 2(5): pp201-207, 2003
6. Adams PD *et al*, Recent developments in the PHENIX software for automated crystallographic structure determination, *J Synchrotron Radiat*, 11(Pt 1): pp53-5, 2004
7. Hendrickson WA, Synchrotron Crystallography, *Trends Biochem Sci*, 25: pp637-643, 2000
8. Morris R *et al*, Breaking good resolutions with ARP/wARP, *J Synchrotron Rad*, 11(1): pp56-59, 2004
9. Hui R and Edwards A, High-throughput protein crystallization, *J Struct Biol*, 142(1): pp154-161, 2003
10. Edwards AM *et al*, Protein production: feeding the crystallographers and NMR spectroscopists, *Nat Struct Biol Suppl*, 7: pp970-972, 2000
11. Waldo GS *et al*, Rapid protein-folding assay using green fluorescent protein, *Nature Biotechnol*, 17: pp691-695, 1999
12. Dale GE, Oefner C and D'Arcy A, The protein as a variable in protein crystallization, *J Struct Biol*, 142(1): pp88-97, 2003
13. Kriminski S *et al*, Flash-cooling and annealing of protein crystals, *Acta Crystallogr*, D58: pp459-471, 2002
14. Garman E and Nave C, Radiation damage to crystalline biological molecules: current view, *J Synchrotron Rad*, 9: pp327-328, 2002
15. Kim Y *et al*, Automation of protein purification for structural genomics, *Journal of Structural and Functional Genomics*, 5(1-2): pp111-118, 2004
16. Page R *et al*, Shotgun crystallization strategy for structural genomics: an optimised two-tiered crystallization screen against the Thermotoga maritima proteome, *Acta Crystallogr*, D59(6): pp1,028-1,037, 2003
17. DiDonato M *et al*, A scaleable and integrated crystallization pipeline applied to mining the Thermotoga maritima proteome, *Journal of Structural and Functional Genomics*, 5(1-2): pp133-146, 2004
18. Segelke BW, Efficiency analysis of sampling protocols used in protein crystallization screening, *J Crystal Growth*, 232: pp553-562, 2001
19. Koppal T, Advancing *in vitro* ADME/Tox, *Drug Discovery and Development*, 5(4): pp47-50, 2004
20. McPherson A, Protein crystallization in the structural genomics era, *Journal of Structural and Functional Genomics*, 5(1-2): pp3-12, 2004
21. Rupp B and Wang J, Predictive models for protein crystallization, *Methods*, 34: pp390-407, 2004
22. Patterson SD, Data analysis – the Achilles heel of proteomics, *Nature Biotechnol*, 21: pp221-222, 2003
23. Keating S, LIMS grows from homemade to custom-made solutions, *Drug Discovery and Development*, 5(4): pp53-54, 2004
24. Haebel PW *et al*, LISA: an intranet-based flexible database for protein crystallography project management, *Acta Crystallogr*, D57(9): pp1,341-1,343, 2001
25. Harris M and Jones TA, Xtrack – a web-based crystallographic notebook, *Acta Crystallogr*, D58(10 Part 2): pp1,889-1,891, 2002
26. Manjasetty BA *et al*, Secure web book to store structural genomics research data, *J Struct Funct Genomics*, 4(2-3): pp121-127, 2003
27. Goh CS *et al*, SPINE 2: a system for collaborative structural proteomics within a federated database framework, *Nucleic Acids Res*, 31(11): pp2,833-8, 2003
28. Pavlis R, Top five myths about LIMS interfacing, *Sci. Computing Instrumentation*, 21(6): pp18-19, 2004
29. Goh CS *et al*, Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis, *J Mol Biol*, 336(1): pp115-30, 2004
30. Bard J and Rhee S, Ontologies in biology: design, applications and future challenges, *Nat Biotechnol*, 5(3): pp213-222, 2004
31. Weselak M *et al*, Robotics for Automated Crystal Formation and Analysis, *Methods Enzymol*, 368: pp45-76, 2003
32. Thorne RE *et al*, Microfabricated mounts for high-throughput macromolecular cryocrystallography, *J Appl Crystallogr*, 36(6): pp1,455-1,460, 2003
33. Sanjo A and Cacheu RE, New Microfabricated Device Technologies for High Throughput and High Quality Protein Crystallization, ICCBM9 Book of Abstracts, pO.I.2, 2002
34. Karain WI *et al*, Automated mounting, centering and screening of crystals for high-throughput protein crystallography, *Acta Crystallogr*, D58: pp1,519-22, 2002
35. Muchmore SW *et al*, Automated crystal mounting and data collection for protein crystallography, *Structure*, 8(12): pp243-246, 2000
36. Snell G *et al*, Automatic Sample Mounting and Alignment System for Biological Crystallography at a Synchrotron Source, *Structure*, 12: pp1-12, 2004
37. Rupp B *et al*, The TB structural genomics consortium crystallization facility: towards automation from protein to electron density, *Acta Crystallogr*, D58: pp1,514-1,518, 2002
38. Jacquamet L *et al*, A new highly integrated sample environment for protein crystallography, *Acta Crystallogr*, D60(5): pp888-894, 2004